Genetic Epidemiology 35:57–69 (2011)

Bayesian Analysis of Rare Variants in Genetic Association Studies

Nengjun Yi* and Degui Zhi

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama

Recent advances in next-generation sequencing technologies facilitate the detection of rare variants, making it possible to uncover the roles of rare variants in complex diseases. As any single rare variants contain little variation, association analysis of rare variants requires statistical methods that can effectively combine the information across variants and estimate their overall effect. In this study, we propose a novel Bayesian generalized linear model for analyzing multiple rare variants within a gene or genomic region in genetic association studies. Our model can deal with complicated situations that have not been fully addressed by existing methods, including issues of disparate effects and nonfunctional variants. Our method jointly models the overall effect and the weights of multiple rare variants and estimates them from the data. This approach produces different weights to different variants based on their contributions to the phenotype, yielding an effective summary of the information across variants. We evaluate the proposed method and compare its performance to existing methods on extensive simulated data. The results show that the proposed method performs well under all situations and is more powerful than existing approaches. *Genet. Epidemiol.* 35:57–69, 2011. © 2010 Wiley-Liss, Inc.

Key words: Bayesian analysis; complex diseases; disparate effects; genetic association; hierarchical models; rare variants; sequence data

Contract grant sponsor: National Institutes of Health; Contract grant numbers: 2R01GM069430-06; GM077490; R00 RR024163.
*Correspondence to: Nengjun Yi, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294-0022.
E-mail: nyi@ms.soph.uab.edu
Received 7 September 2010; Revised 15 October 2010; Accepted 8 November 2010
Published online 10 December 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/gepi.20554

INTRODUCTION

It has been a well-established hypothesis that the genetic etiology of common (or complex) human diseases is determined by both common and rare genetic variants [Bodmer and Bonilla, 2008; Schork et al., 2009]. Although genome-wide association studies, which have thus far focused on common variants (with minor allele frequency (MAF) $>\sim$5%) in the human genome, have successfully identified hundreds of novel disease-associated variants, these common variants explain only a small proportion of heritability for most diseases, motivating interest in finding the 'missing heritability' [Eichler et al., 2010; Manolio et al., 2009]. Rare variants have been naturally speculated as one of the most important sources of missing heritability [Cirulli and Goldstein, 2010; Eichler et al., 2010; Manolio et al., 2009]. Several studies have already shown that rare variants play an important role in genetic determination for some diseases [Ahituv et al., 2007; Azzopardi et al., 2008; Cohen et al., 2004, 2006; Ji et al., 2008; Nejentsev et al., 2009; Romeo et al., 2007, 2009]. Recent advances in next-generation sequencing technologies facilitate the detection of rare variants, making it possible to uncover the roles of rare variants in complex diseases.

As a single rare variant contains little variation owing to low MAF ($<$0.5 or 1%), statistical methods that test variants individually provide insufficient power to detect causal rare variants. Therefore, association analysis of rare variants requires sophisticated methods that can effectively combine the information across variants and test for their overall effect [Manolio et al., 2009]. Several approaches have been developed to analyze rare variants, including the Collapsing, Simple-Sum, and Weighted-Sum methods [Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Price et al., 2010]. These methods summarize multiple rare variants by weighting them equally [Li and Leal, 2008; Morris and Zeggini, 2010] or on the basis of estimated standard deviation [Madsen and Browning, 2009] or functional prediction [Price et al., 2010]. As we show in this study, however, these existing methods actually implicate assumptions about the relative effect sizes of individual variants (for example, the Simple-Sum method implicitly assumes that the genetic effects of individual variants are identical) and thus can be sub-optimal if the data do not follow the implicit assumptions.

There are complications that the existing methods have not addressed yet. First, multiple rare variants detected in a gene or region may affect phenotype in either direction (i.e. some are disease-causing and others are disease-protective) [Cohen et al., 2004; Manolio et al., 2009]. If these rare variants are simply pooled, the existing methods will fail, because the effects of the variants can cancel each other and thus the true signal is lessened. Second, sequencing uncovers both functional and nonfunctional variants, and treating them equally may reduce association. The ideal solution to these problems is to separately combine disease-causing and protective variants or scale the allele counts of all variants in the same association direction and to exclude nonfunctional variants from the analysis. However, accurately determining which variants

are disease-causing or protective and which are responsible for a given phenotype represent a massive task and are not always feasible [Manolio et al., 2009]. Therefore, statistical methods that can deal with these complications are required.

In this study, we introduce a novel Bayesian hierarchical generalized linear model for analyzing multiple rare variants within a gene or genomic region in association studies. Although our method can deal with various phenotypes, we demonstrate its performance with a binary disease trait as in population-based case-control studies. Rather than predetermining the weights of variants as previous methods, our approach jointly models the overall effect and the weights of multiple rare variants and estimates them from the data. This could produce different weights to different variants based on their contributions to the phenotype, yielding an effective summary of the information across variants. We use extensive simulations to evaluate the proposed method and compare its performance to existing methods. The results show that the proposed method performs well under all situations and is more powerful than existing approaches.

METHODS

BAYESIAN MODEL OF MULTIPLE RARE VARIANTS

Suppose that a population-based association study consists of n unrelated individuals, phenotyped for a binary disease trait y_i (i.e. if diseased, $y_i = 1$; otherwise, $y_i = 0$), and genotyped for m rare variants in a candidate gene or functional genomic region. We denote the genotypes of variant j by A_jA_j, A_ja_j, or a_ja_j, where a_j is the minor allele with the observed frequency $p_j < 1\%$. The relation between the disease status and the genotypes of m rare variants can be expressed by a generalized linear regression

$$h(\Pr(y_i = 1)) = \mu + \sum_{j=1}^{m} x_{ij}\beta_j, \quad (1)$$

where the link function h is the logit or probit function, μ is the intercept, β_j is the main effect for the jth variant, and x_{ij} is the main-effect predictor for the ith individual at the jth variant. For an additive model, $x_{ij} = 0$, 1, or 2 for A_jA_j, A_ja_j, or a_ja_j, and for a dominant model, $x_{ij} = 0$ or 1 for A_jA_j or A_ja_j and a_ja_j, respectively. For a rare variant, the additive model is approximately equivalent to the dominant model because the frequency of a_ja_j is extremely low.

The association between the disease and the variants may be examined by testing $\beta_j = 0$, $j = 1,\ldots,m$. For rare variants, however, such an analysis is underpowered because a single variant explains very low genetic variation and typically is undetectable. Under the additive model and Hardy-Weinberg equilibrium, for example, the genetic variance of the jth variant is $\mathrm{Var}(x_{ij}\beta_j) = 2p_j(1-p_j)\beta_j^2$, equal to 9.5×10^{-3} when the frequency p_j and the odds ratio $\exp(\beta_j)$ equal 1 and 2%, respectively. A solution would be to create a "genetic score" that combines information across multiple rare variants for each individual. The genetic score is then treated as a single predictor, allowing us to detect the overall association of the variants with the disease. We construct the genetic score as a linear function of the separate main-effect predictors, i.e. $T_i = \sum_{j=1}^{m} \alpha_j x_{ij}$, and set up a generalized linear model:

$$h(\Pr(y_i = 1)) = \mu + \left(\sum_{j=1}^{m} \alpha_j x_{ij}\right)\beta. \quad (2)$$

In this model, the common coefficient β represents the *overall effect* for the m rare variants, and the α_j's can be interpreted as the *relative effects* or *weights* of the individual variants. To investigate the overall association, we test the hypothesis $\beta = 0$.

Instead of presetting α_j's as in existing methods, it would be better to estimate them from the data. But we cannot simply use classical framework (i.e. setting uniform distributions on the α_j's), since this would result in a nonidentifiable model and thus be equivalent to estimating a separate coefficient for each of the variants [Gelman, 2004; Gelman and Hill, 2007]. However, we can set up an informative prior for the α_j's, so that the model is identifiable. We use the Student-t distribution for the α_j's:

$$\alpha_j \sim N(\mu_j, \tau_j^2), \quad \tau_j^2 \sim \mathrm{Inv}{-}\chi^2(1, s_\alpha^2) \quad \text{for } j = 1,\ldots,m, \quad (3)$$

with the scale s_α set to a low value such as 0.5 [Gelman et al., 2008]. This prior distribution constrains α_j's to be fairly close to the prior mean μ_j's, but allows for different values. An alternative prior is to use the normal distribution with a fixed variance, i.e. $\tau_j^2 = 0.3$ [Gelman and Hill, 2007]. However, we prefer the Student-t distribution because it estimates the variances τ_j^2 from the data and thus may better deal with disparate effects. The prior means μ_j can be specified as the relative importance of the individual variants based on our prior knowledge or initial analysis (see Discussion). In this study, we incorporate no prior information into the model by setting $\mu_j = 1$ for all $j = 1,\ldots,m$. We found that the method is fairly robust to any small changes for the scale s_α (for example, from 0.2 to 0.8).

The common coefficient β usually can be estimated classically. However, low allelic frequencies can yield very small variance $\mathrm{Var}(T_i)$, for which the classical procedure often results in numerically instable estimate. To overcome this problem, we use a weakly informative prior to constrain β in a reasonable range. Following Gelman et al. [2008], we place a Student-t distribution with center 0, degree of freedom 1, and scale 2.5 on β:

$$\beta \sim N(0, \tau_\beta^2), \quad \tau_\beta^2 \sim \mathrm{Inv}{-}\chi^2(1, 2.5^2). \quad (4)$$

COMPUTATION

Our Bayesian generalized linear model can be fitted using Markov chain Monte Carlo algorithms that fully explore the joint posterior distribution of the parameters by alternatively sampling each parameter from its conditional posterior distribution. However, it is desirable to have a faster computation that provides a point estimate (i.e. the posterior mode) of β and α_j's and their standard errors (and thus the P-values) by maximizing the marginal posterior $p(\beta, \alpha_1, \ldots, \alpha_m, \mu | y, X)$. Such an approximate calculation has been routinely applied in statistical practice [Gelman et al., 2008]. We develop our algorithm by modifying the standard iterative weighted least squares (IWLS) for fitting classical generalized linear models. We have implemented these computations by altering the

`glm` function in R (the general statistical package) that fits classical generalized linear models.

Our algorithm simultaneously estimates the parameters α_j's and β using an iterative procedure. We initialize the algorithm by setting β to the value estimated from the Simple-Sum method or setting $\alpha_1 = \cdots = \alpha_m = 1$. Then, at each step of the algorithm, we first update α_j's conditional on the current estimate $\hat{\beta}$ by using the modified IWLS algorithm of Yi and Banerjee [2009] and Yi et al. [2010] to fit the hierarchical generalized linear model:

$$h(\Pr(y_i = 1)) = \mu + \sum_{j=1}^{m} (x_{ij}\hat{\beta})\alpha_j,$$
$$\alpha_j \sim N(\mu_j, \tau_j^2), \quad \tau_j^2 \sim \text{Inv}{-}\chi^2(1, s_\alpha^2) \quad \text{for } j = 1, \ldots, m. \quad (5)$$

We then update β conditional on the current estimates $\hat{\alpha}_j$'s by fitting the hierarchical model using the modified IWLS algorithm [Yi and Banerjee, 2009; Yi et al., 2010]:

$$h(\Pr(y_i = 1)) = \mu + \left(\sum_{j=1}^{m} \hat{\alpha}_j x_{ij}\right)\beta,$$
$$\beta \sim N(0, \tau_\beta^2), \quad \tau_\beta^2 \sim \text{Inv}{-}\chi^2(1, 2.5^2). \quad (6)$$

Instead of doing a full IWLS for each of these two models, we can perform one step of weighted least squares at each iteration, thus taking less computer time to ultimately achieve convergence by not wasting time getting hyper-precise estimates at each step of the algorithm. We apply the criterion in the `glm` function to assess convergence, i.e.

$$\frac{|d_1^{(t)} - d_1^{(t-1)}|}{0.1 + |d_1^{(t)}|} < \varepsilon \quad \text{and} \quad \frac{|d_2^{(t)} - d_2^{(t-1)}|}{0.1 + |d_2^{(t)}|} < \varepsilon,$$

where $d_1^{(t)}$ and $d_2^{(t)}$ are deviances at the tth iteration for the models [4] and [5], respectively, and ε is a small value (say 10^{-8}). In practice, our algorithm converges rapidly. At convergence of the algorithm, we obtain all the outputs produced by the `glm` function, including the latest estimate $\hat{\beta}$, their standard deviations, and the P-values for testing $\beta = 0$.

RELATIONSHIP WITH EXISTING METHODS

The basic procedure of rare variant analysis is to construct a weighted combination (genetic score) of m rare variants, $T_i = \sum_{j=1}^{m} \alpha_j x_{ij}$, that summarizes the information across the variants for each individual i, and then estimate the association between the phenotype y_i and the genetic score T_i using a generalized linear model, $h(\Pr(y_i = 1)) = \mu + T_i\beta$, or other testing statistics. Our method differs from existing methods in estimating the weights α_j's (along with the overall effect β) from the data using a hierarchical modeling framework rather than simply presetting them to fixed values. This would produce higher weights for more 'important' variants.

Presetting the weights α_j's to different values results in different existing methods: (1) If $\alpha_1 = \cdots \alpha_m = 1$, we have $T_i = \sum_{j=1}^{m} x_{ij}$ and thus the method is the Simple-Sum [Han and Pan, 2010; Morris and Zeggini, 2010]; (2) If we take $T_i = I(\sum_{j=1}^{m} x_{ij})$, where $I(x)$ is an indicator variable taking 1 if $x > 0$, and 0 otherwise, the method becomes the Collapsing approach [Li and Leal, 2008]; (3) If $\alpha_j = 1/\text{sd}(x_{ij} \mid y_i = 0)$, $j = 1,\ldots,m$, where $\text{sd}(x_{ij} \mid y_i = 0)$ is the estimated standard deviation of x_{ij} in unaffected individuals, the model is similar to the Weighted-Sum approach [Madsen and Browning, 2009]; (4) If setting α_j to the posterior probability of being functional for each variant j, T_i corresponds to that of Price et al. [2010]. These posterior probabilities can be calculated using bioinformatics tools such as PolyPhen [Adzhubei et al., 2010; Price et al., 2010].

From the above procedure, we can see that the term $\alpha_j\beta$ actually corresponds to the genetic effect β_j of the jth variant. With the fixed weights, therefore, the individual genetic effect β_j is proportional to the corresponding weight α_j. This important result reveals the underlying assumptions of the existing methods. The Simple-Sum method implicates that the effects of all variants are identical; obviously, this is an unrealistic assumption. The Weighted-Sum method first standardizes the main-effect predictors and then assumes identical coefficients for all variants in the standardized model. This corresponds to the implicit assumption $\beta_j \propto 1/\text{sd}(x_{ij} | y_i = 0)$. The approach of Price et al. [2010] implicitly assumes that the effect of a variant is proportional to the posterior functional probability. Therefore, all the existing methods implement pooling of multiple rare variants according to certain assumptions about the genetic effects of variants. Although powerful in certain situations, these methods can be inefficient if the underlying assumption is not true. In contrast, the proposed method does not require any assumptions about the relative importance of individual variants and thus could be more robust than the existing methods.

SIMULATIONS AND COMPARISON WITH EXISTING METHODS

We use extensive simulations to evaluate the proposed approach and to compare the proposed method with five existing methods: the Collapsing, Simple-Sum, Weighted-Sum, and All-Variants (i.e. jointly fitting all variants) and Single-Variant (i.e. fitting one variant at a time).

BAYESIAN VERSIONS OF EXISTING METHODS

Although various testing statistics have been proposed for the existing Collapsing, Simple-Sum, and Weighted-Sum methods [Han and Pan, 2010; Li and Leal, 2008; Madsen and Browning, 2009; Price et al., 2010], we implement these methods using logistic regressions:

$$\text{logit}(\Pr(y_i = 1)) = \mu + T_i\beta,$$

with $T_i = I(\sum_{j=1}^{m} x_{ij})$ for the Collapsing, $T_i = \sum_{j=1}^{m} x_{ij}$ for the Simple-Sum, and $T_i = \sum_{j=1}^{m} x_{ij}/\text{sd}(x_{ij}|y_i = 0)$ for the Weighted-Sum. The All-Variants method jointly estimates the individual effects of all variants:

$$\text{logit}(\Pr(y_i = 1)) = \mu + \sum_{j=1}^{m} x_{ij}\beta_j,$$

and the Single-Variant separately estimates the effects of individual variants:

$$\text{logit}(\Pr(y_i = 1)) = \mu + x_{ij}\beta_j, \quad j = 1, \ldots, m.$$

These logistic regressions can be nonidentifiable when the variance $\mathrm{Var}(T_i)$ or $\mathrm{Var}(x_{ij})$ is small [Li and Leal, 2008]. We overcome this problem by placing the weakly informative prior [Gelman et al., 2008], $\beta \sim N(0, \tau_\beta^2)$, $\beta_j \sim N(0, \tau_\beta^2)$, $j = 1, \ldots, m$, and $\tau_\beta^2 \sim \mathrm{Inv}{-}\chi^2(1, 2.5^2)$. We fit these models using the modified IWLS algorithm of Yi and Banerjee [2009] and Yi et al. [2010]. This improves the performance of these previous approaches and has the advantage of always producing stable estimates.

SIMULATION DESIGN

We consider different combinations of the factors that may affect the performance of the methods:

(a) *Sample size*: We simulate $n = 500$, 1,000, and 2,000 individuals with an equal number of affected and unaffected.

(b) *Number of rare variants*: We simulate $m = 20$, 40, and 80 rare variants.

(c) *Minor allelic frequencies and genotypes*: We sample m variants independently because correlation between rare variants is low [Pritchard, 2001; Pritchard and Cox, 2002]. For the jth variant, we sample the MAF p_j uniformly from the region [0.001, 0.01], as variants with MAF < 0.001 would be indistinguishable in our presumed sample sizes. Assuming the Hardy-Weinberg equilibrium for each variant, we thus generate the genotypes from the multinomial distribution: $\mathrm{Multin}(n; (1-p_j)^2, 2(1-p_j)p_j, p_j^2)$.

(d) *Genetic model*: We evaluate our method using the additive genetic model for each variant. For rare variants, the additive model is approximately equivalent to the dominant model, and detection of recessive effects requires extremely large sample [Li and Leal, 2008].

(e) *Number of functional variants and genetic effects*: For $m = 20$, 40 and 80, we set all the variants to be functional or randomly sample 40% of the simulated variants as nonfunctional. For each functional variant, we simulate the odds ratio $\exp(\beta_j)$ to be 1 (for type-I error rate) or uniformly from the region $[1.05, \mathrm{OR_u}]$ (for power analysis). To ensure that the overall effect of all variants is reasonably low, we determine the upper bound $\mathrm{OR_u}$ by controlling the total liability heritability, which approximates $h_T^2 = 2p_{\mathrm{ave}}(1-p_{\mathrm{ave}})\beta_{\mathrm{ave}}^2 m_f / [2p_{\mathrm{ave}}(1-p_{\mathrm{ave}})\beta_{\mathrm{ave}}^2 m_f + 1.6^2]$ (as derived in the next paragraph), where p_{ave} is the average MAF, m_f is the number of functional variants, and $\exp(\beta_{\mathrm{ave}})$ is the average odds ratio, equal to $(1.05+\mathrm{OR_u})/2$. For example, $\mathrm{OR_u} = 2.0$ when $m_f = 40$ and $h_T^2 = 3\%$. We consider the total liability heritability h_T^2 from 0.7 to 8%. Finally, we consider the most complicated case in which the effects of the functional rare variants are in opposite directions; for each functional variant, we first simulate the odds ratio $\exp(\beta_j)$ uniformly from the region $[1.05, \mathrm{OR_u}]$, and then change the sign of β_j with the probability of 0.3 or 0.5.

Given the coefficients β_j and the genotypic codes x_{ij}, we can simulate the disease phenotype y_i using two methods. The first method is to directly sample y_i from the binomial distribution: $\mathrm{Bin}(1, \mathrm{logit}^{-1}(\sum_{j=1}^m x_{ij}\beta_j))$. This procedure is repeated until we obtain $n/2$ affected and $n/2$ unaffected individuals. The second is to use the latent-data formulation of the logistic regression; the logistic model $\mathrm{logit}(y_i = 1) = \sum_{j=1}^m x_{ij}\beta_j$ is equivalent to the model, $w_i \sim N(\sum_{j=1}^m x_{ij}\beta_j, 1.6^2)$, $y_i = 1 \Leftrightarrow w_i > c$ [see Gelman and Hill, 2007]. Thus, we first sample n latent normal phenotypes w_i and then set $n/2$ individuals with the 50% largest w_i as affected (i.e. $y_i = 1$) and the other $n/2$ individuals as unaffected. The latent-data formulation allows us to calculate the proportion of the latent-data variance explained by the variants, i.e. the liability heritability [Wray et al., 2010], $h^2 = \sum_{j=1}^m 2p_j(1-p_j)\beta_j^2 / [\sum_{j=1}^m 2p_j(1-p_j)\ \beta_j^2 + 1.6^2]$. As described above, this formulation can be used to control the total heritability when we simulate the coefficients β_j.

For each set of parameters, 1,000 replicated data sets are simulated, and each is analyzed using our hierarchical model approach and the Collapsing, Simple-Sum, Weighted-Sum, All-Variants, and Single-Variant methods. For each analysis, we use the additive genetic model. We calculate power to detect overall association $\beta = 0$ at significance levels of $\alpha = 0.001$ and $\alpha = 2.5 \times 10^{-6}$. These thresholds correspond to candidate gene studies [Price et al., 2010] or a genome-wide study of about 20,000 fairly independence human genes [Madsen and Browning, 2009], respectively. We also examine type-I error rate at a significance level of $\alpha = 0.05$. For the All-Variants and Single-Variant methods, the overall association is examined by testing whether at least one $\beta_j = 0$ for all $j = 1, \ldots, m$, and for simplicity we do not adjust the significance level for multiple testing. Therefore, we overestimate the power for the All-Variants and Single-Variant methods.

RESULTS

TYPE-I ERROR RATE

As shown in Figure 1, the type-I error rates are well controlled for the proposed method and the Collapsing,

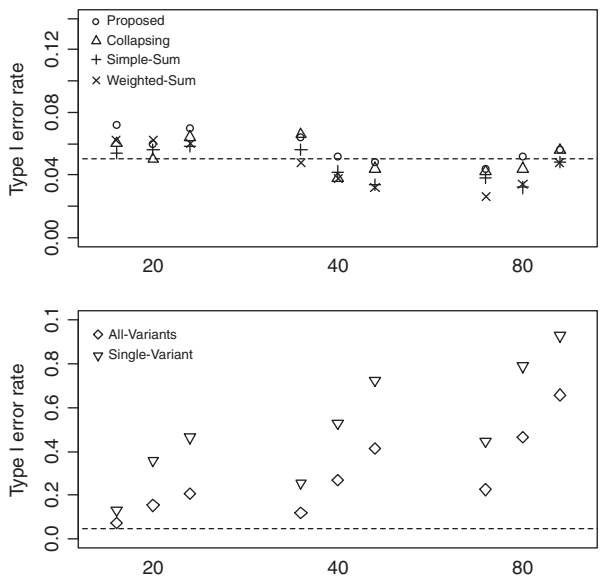

Fig. 1. Type-I error rates of the proposed method, Collapsing, Simple-Sum, Weighted-Sum, All-Variants, and Single-Variant methods at the 5% level with the number of variants $m = 20$, 40, and 80 and the number of individuals $n = 500$, 1,000, and 2,000. The dashed horizontal line is the nominal 0.05 level.

Simple-Sum, and Weighted-Sum methods. The proposed method slightly inflates the type-I error rate when $n = 500$ and $m = 20$. However, there is no constant trend that the proposed method generates higher type-I error rate than those of the previous methods. The type-I error rates for the All-Variants and Single-Variant methods are unacceptably high, and significantly increases with increasing number of variants, indicating the need for multiple-testing correction.

ANALYSIS OF FUNCTIONAL VARIANTS

We first investigated powers of the methods for the relatively simple scenario in which all rare variants are functional and affect disease risk in the same direction. As shown in Figures 2 and 3, the results for different sample sizes, different numbers of rare variants and different liability heritabilities display similar patterns of empirical powers. A notable result is that the proposed method is consistently more powerful than the other methods. Because our method estimates the weights of multiple variants from the data, our model fits the data better and generates a genetic score that better summarizes the information across the variants. Thus, the proposed method improves the power to detect the overall association between rare variants and disease.

As expected, the power drastically increases with the sample size and the total liability heritability explained by the variants. These relationships hold rather generally for the methods that we examined. Given sample size and total liability heritability, the power slightly decreases with the number of variants. This is likely the results

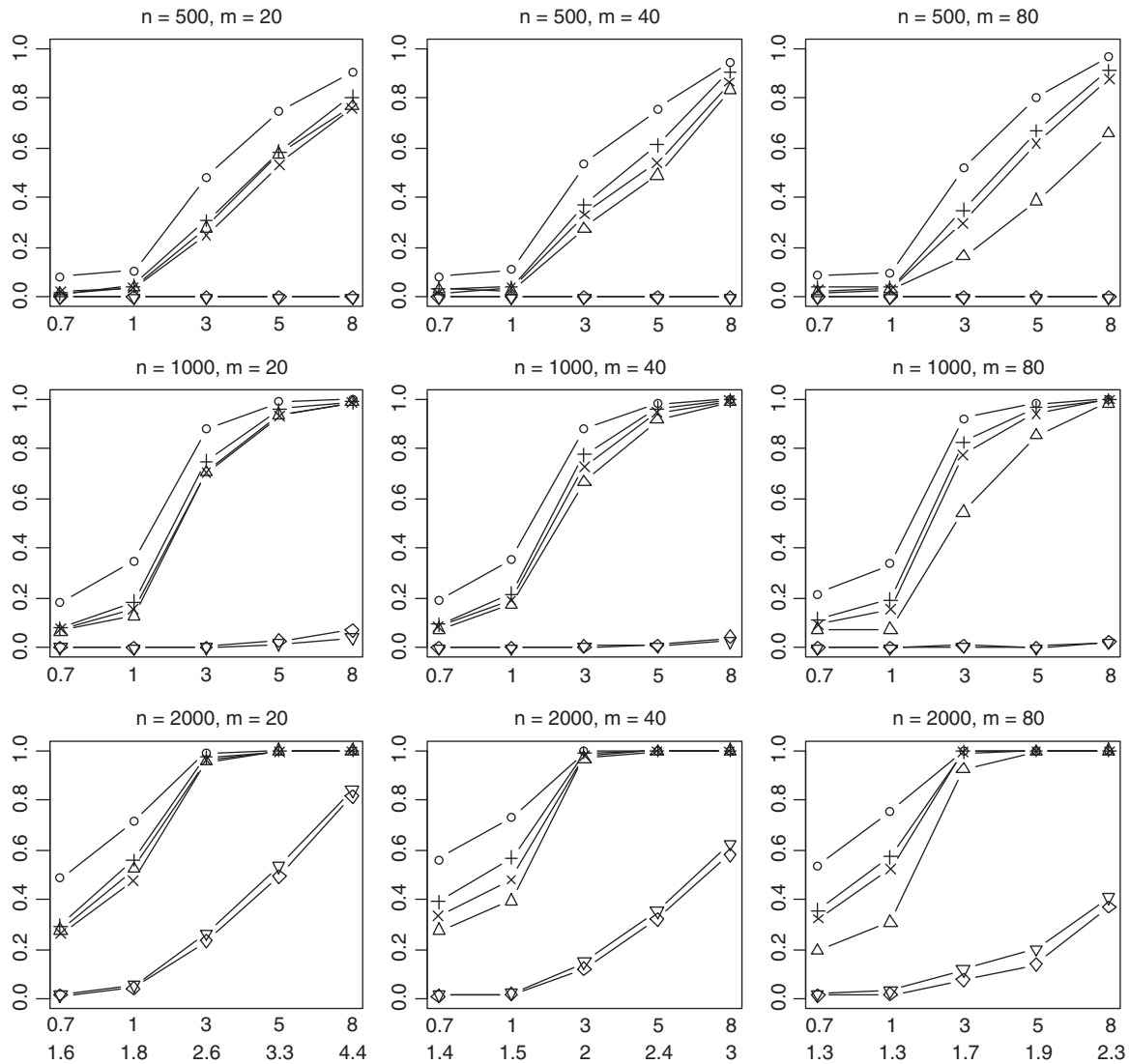

Fig. 2. Empirical powers of the proposed method (○), Collapsing (△), Simple-Sum (+), Weighted-Sum (×), All-Variants (◇), and Single-Variant (▽) methods at a significance level of $\alpha = 0.001$. n and m represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. The numbers at the bottom line are the corresponding upper bounds OR_u of the odds ratios. All the simulated variants are functional, and affect phenotype in the same direction.

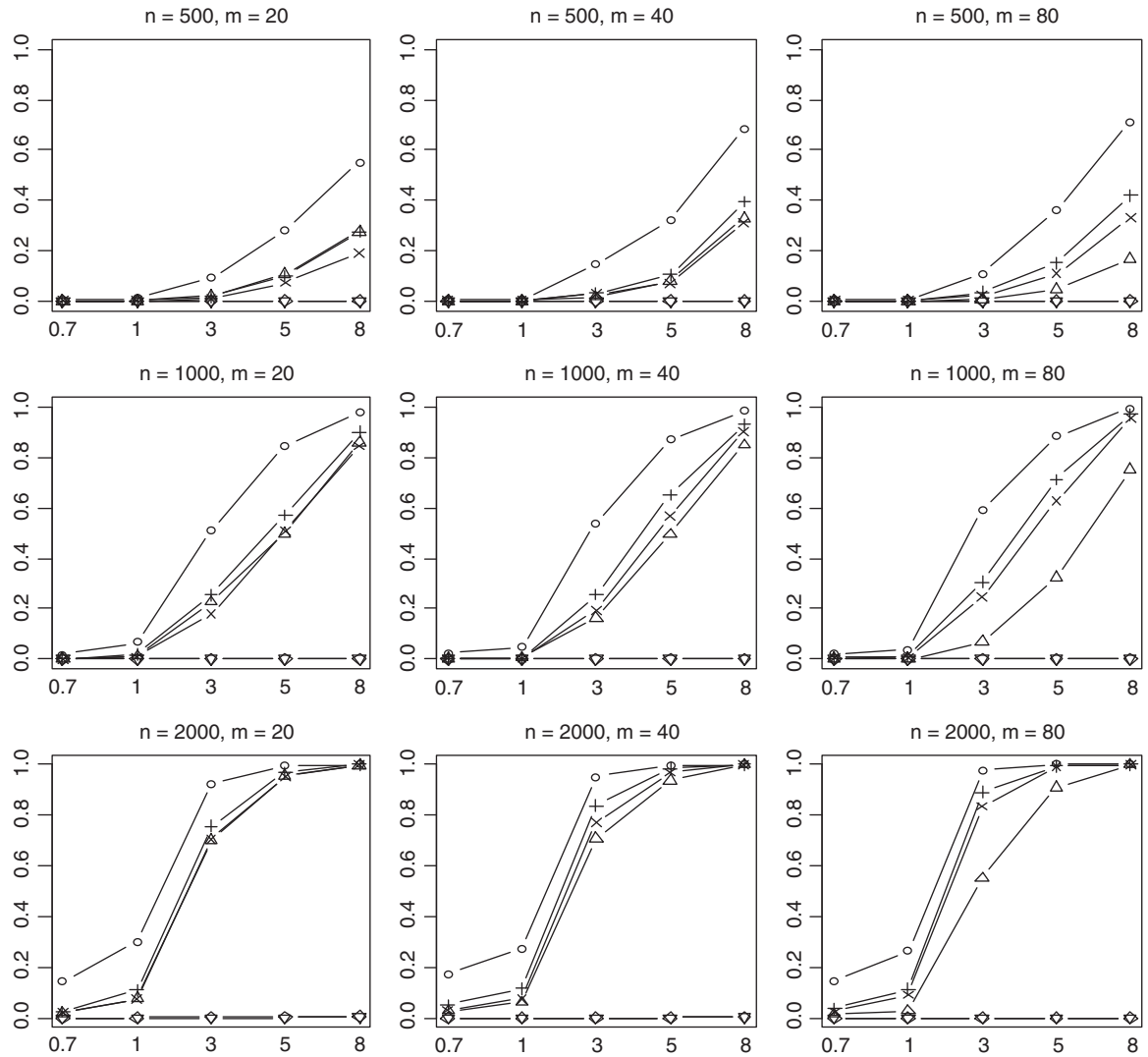

Fig. 3. Empirical powers of the proposed method (○), Collapsing (△), Simple-Sum (+), Weighted-Sum (×), All-Variants (◇), and Single-Variant (∇) methods at a significance level of $\alpha = 2.5 \times 10^{-6}$. *n* and *m* represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. All the simulated variants are functional, and affect phenotype in the same direction.

that more variants generate a lower upper bound OR_u of odds ratios for individual effects (thus smaller individual effects) and thus their information may be more difficult to be summarized. Our simulations showed that with small sample sizes ($n = 500$, 1,000) the All-Variants and Single-Variant methods have no power to detect the association between rare variants and disease. These are expected because these methods test for the effects of single variants each of which has little variation. For larger sample size ($n = 2{,}000$), the powers of the All-Variants and Single-Variant methods go up at the significance level of $\alpha = 0.001$ with no multiple-testing correction, but rapidly decrease to near zero with a more stringent significance level (Fig. 3).

Our results showed that some of the previous methods produce similar power as the proposed method in some situations, masking the real difference between these methods. To investigate whether the proposed method provides any advantages in these situations, we calculated the median value of *P*-values for simulation replicates with *P*-value < 0.001. A notable outcome of this analysis is that the proposed method uniformly yields much lower *P*-values than the previous methods (Fig. 4). This finding indicates that our method usually provides stronger evidence of association if the variants really influence the disease.

INCLUSION OF NONFUNCTIONAL VARIANTS

Nonfunctional variants do not contribute to disease risk. Therefore, the inclusion of nonfunctional variants in the analysis introduces noisy variation in the model and may influence the performance of the methods. Our simulations showed that the power decreases when

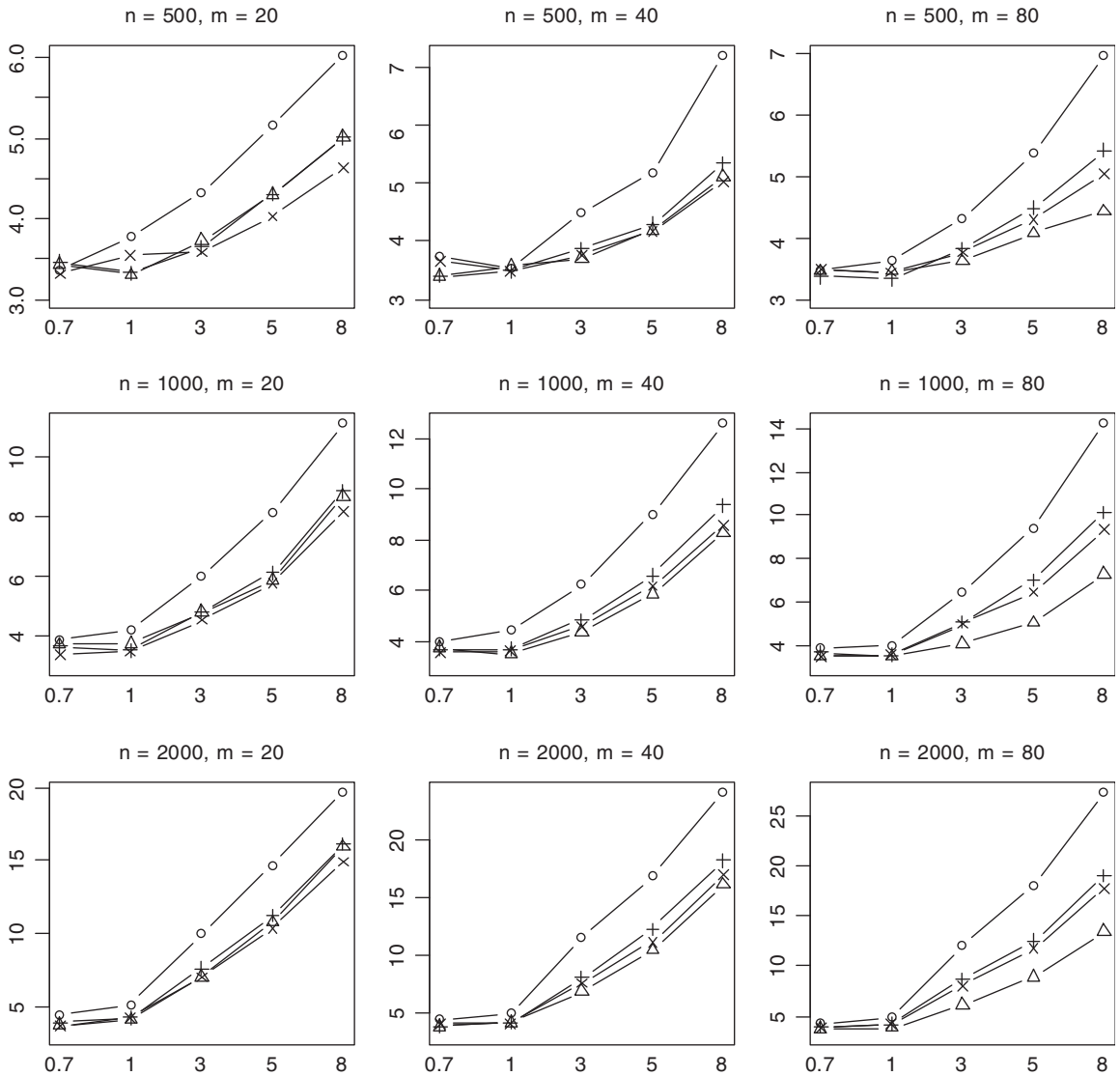

Fig. 4. Median of *P*-values (rescaled as $-\log_{10}P$) for the proposed method (○), Collapsing (△), Simple-Sum (+), and Weighted-Sum (×) methods for replicates with *P*-value <0.001. *n* and *m* represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. All the simulated variants are functional, and affect phenotype in the same direction.

nonfunctional variants are included (Figs. 5 and 6). This is true for all the methods that we examined. However, we found that the previous methods lose more power than the proposed method. This probably results from the fact that the previous methods use equal or inappropriate weights for functional and nonfunctional variants, thereby ineffectively summarizing the information across the multiple rare variants. In contrast, our method estimates weights from the data and thus can set lower or even zero weights to nonfunctional variants, providing a better genetic score.

Although the power decreases with inclusion of nonfunctional variants, the general conclusions obtained earlier still hold. The proposed method is uniformly more powerful (Figs. 5 and 6) and generates much lower *P*-values than the previous methods (Fig. 7).

ANALYSIS OF RARE VARIANTS WITH OPPOSITE EFFECTS

We finally investigated empirical power of the methods in the complicated scenario in which the effects of the functional rare variants influence disease in opposite directions. With 30% (70%) of functional variants increasing (decreasing) disease risk, the previous methods have some power to detect the association when sample size is large ($n = 2{,}000$) (Fig. 8). But the power rapidly decreases with a more stringent significance level (Fig. 9). For the worst case where 50% (50%) of functional variants increase (decrease) disease risk, the Collapsing, Simple-Sum, and Weighted-Sum methods have no power to detect the association even when sample size and odds ratios are large. These results are expected because

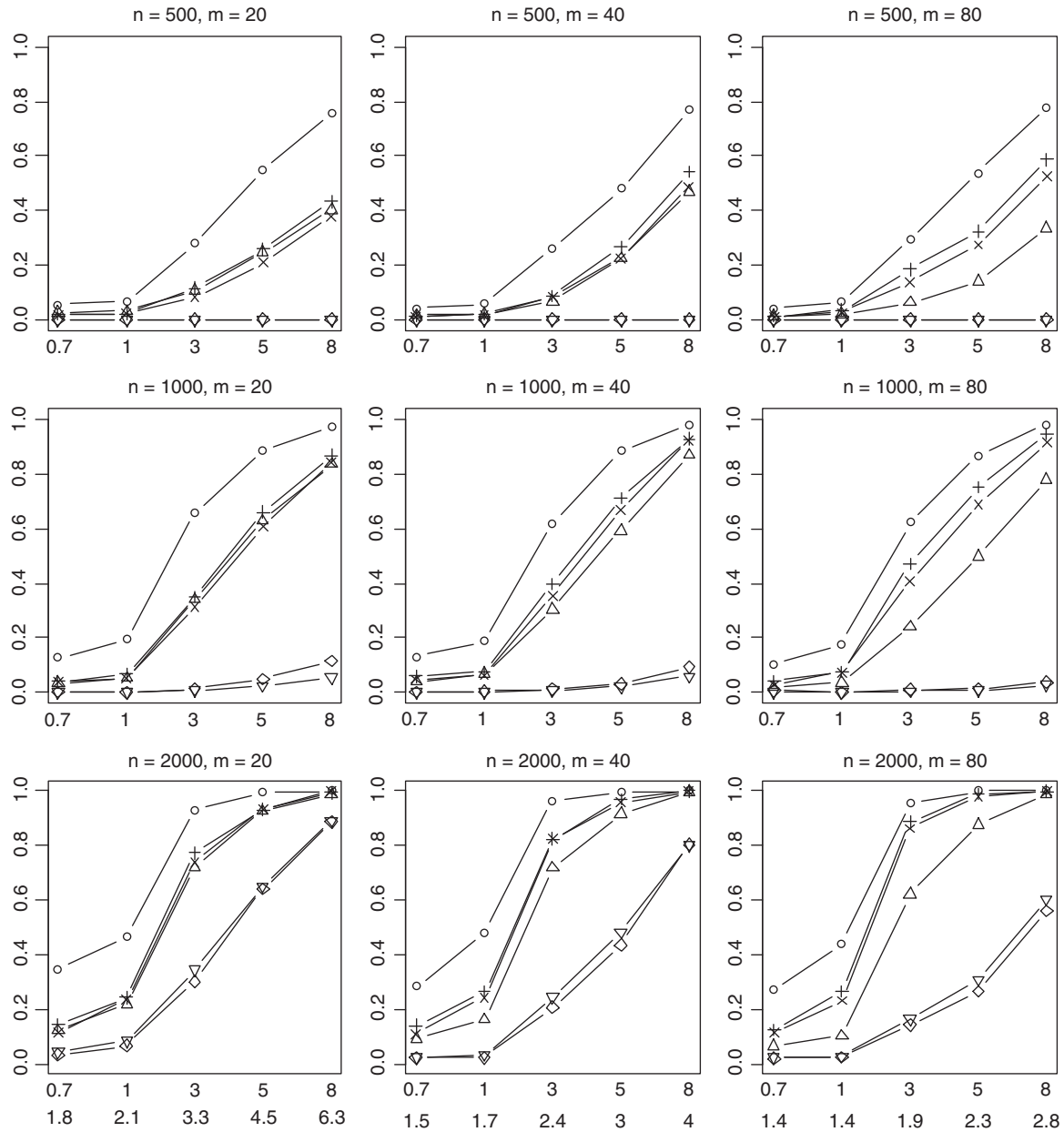

Fig. 5. Empirical powers of the proposed method (○), Collapsing (△), Simple-Sum (+) and Weighted-Sum (×), All-Variants (◇), and Single-Variant (▽) methods at a significance level of $\alpha = 0.001$. n and m represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. The numbers at the bottom line are the corresponding upper bounds OR_u of the odds ratios. Sixty per cent of the simulated variants are functional, and affect phenotype in the same direction.

these methods simply pool all variants together, using equal weights for disease-causing and disease-protective variants. Therefore, the information across multiple rare variants is canceled and the true association signal is completely hidden. As expected, the All-Variants and Single-Variant methods perform similarly as the previous cases that we studied.

The most striking finding of this study is that our method is still powerful even when multiple rare variants have opposite effects on disease risk (Figs. 8 and 9). This remarkable feature is certainly the result of the unique property of the proposed method. Our method estimates weights from the data and thus yields different weights for disease-causing and protective variants, avoiding cancellation of individual-variant variation. Compared to the simpler case, however, for this complicated case, the proposed method is less powerful and is more sensitive to the number of variants. This is expected because the increasing complexity certainly reduces the accuracy of statistical inference.

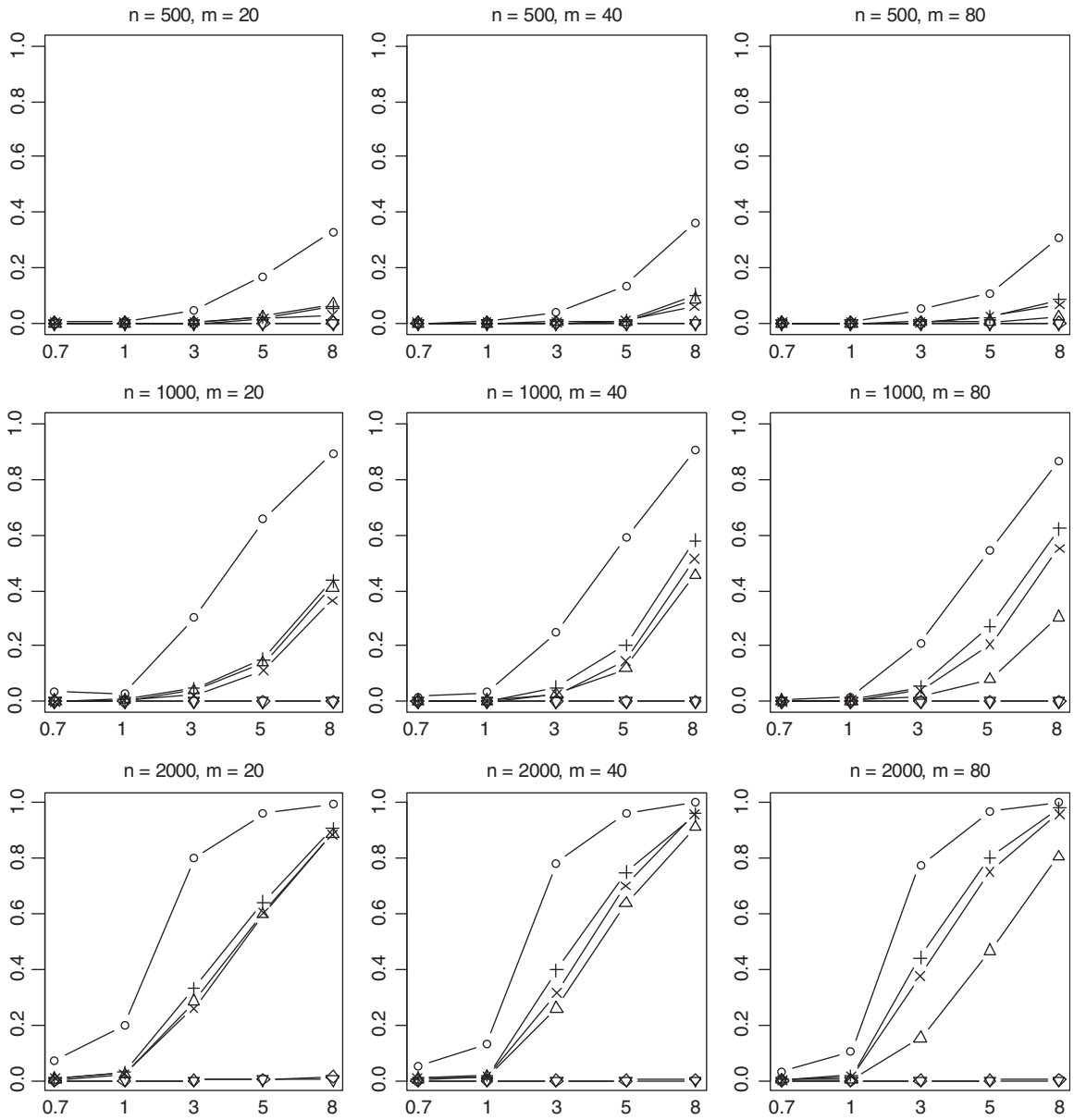

Fig. 6. Empirical powers of the proposed method (○), Collapsing (△), Simple-Sum (+) and Weighted-Sum (×), All-Variants (◇), and Single-Variant (▽) methods at a significance level of $\alpha = 2.5 \times 10^{-6}$. n and m represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. Sixty per cent of the simulated variants are functional, and affect phenotype in the same direction.

DISCUSSION

The Bayesian method developed here includes innovative and attractive features in both modeling and computation steps. The proposed hierarchical model treats the weights as parameters, not only obviating the choice of them but also allowing for better combination of multiple variants. The key to this approach is the use of an appropriate model for the weights, so that the overall coefficients and the weights are identifiable [Gelman, 2004; Gelman and Hill, 2007]. The proposed algorithm extends the standard procedure for fitting classical generalized linear models in the general statistical package R to our Bayesian model, leading to the development of stable and flexible software. Although a fully Bayesian computation that explores the posterior distribution of parameters provides more information, our mode-finding algorithm quickly produces all results as in routine statistical analysis. Our method is directly applicable to candidate gene association studies and has the potential to be applied to large-scale exome sequencing or whole-genome resequencing data. Furthermore, the hierarchical

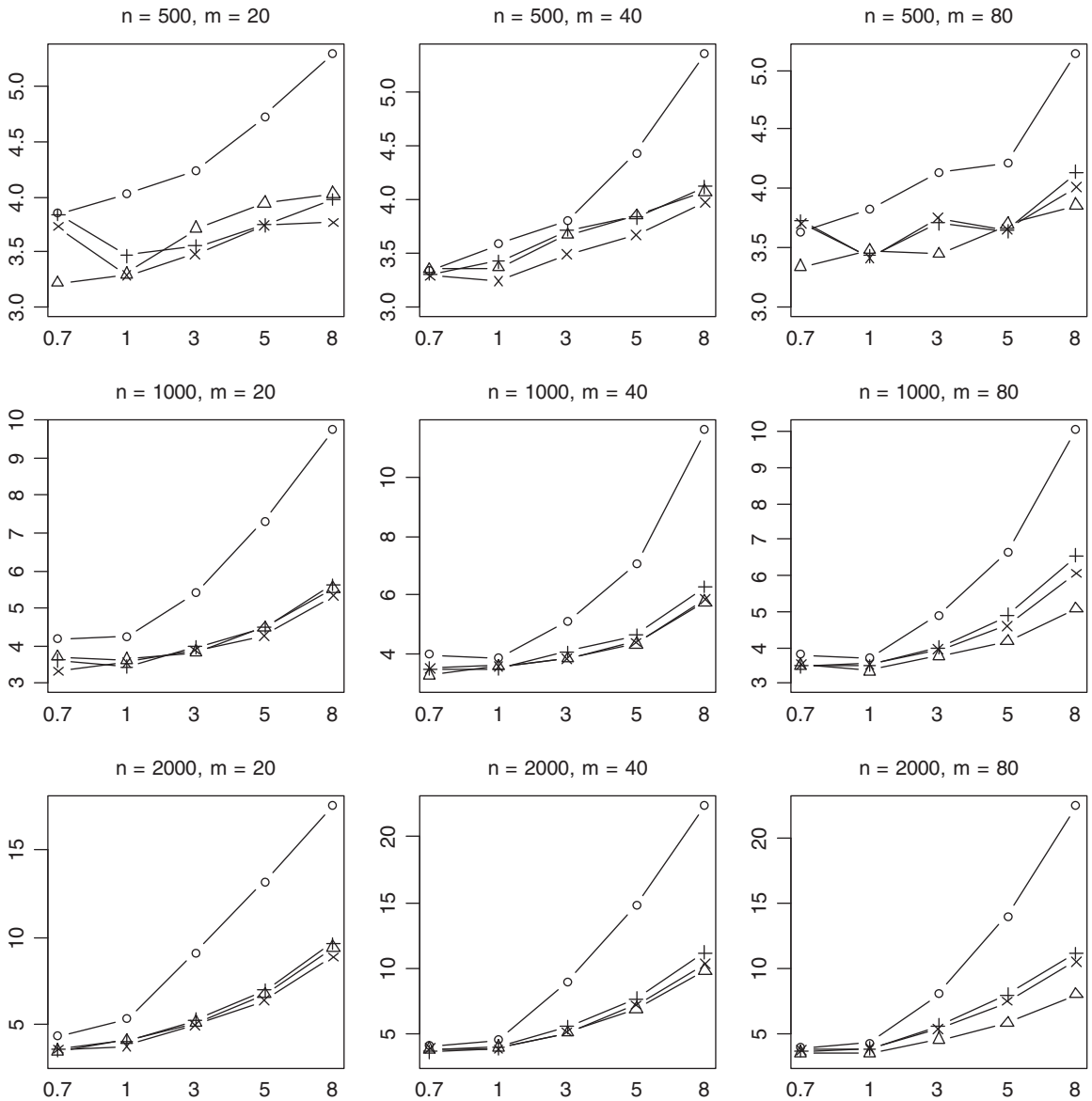

Fig. 7. Median of *P*-values (rescaled as $-\log_{10} P$) for the proposed method (○), Collapsing (△), Simple-Sum (+), and Weighted-Sum (×) methods for replicates with *P*-value <0.001. *n* and *m* represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. Sixty per cent of the simulated variants are functional, and affect phenotype in the same direction.

generalized model framework and the computational strategy developed here can deal with various types of continuous and discrete phenotypes and any generalized models.

We describe our Bayesian method by setting the same prior means for all variants. This means that our model assumes no hypothesis on the relative effect size of rare variants. The motivation for this prior specification is that our understanding of the role of rare variants in common disease is far from complete and thus any assumptions may not be always appropriate. However, recent empirical and theoretical studies have suggested that effect size may correlate with the frequency distribution or the functional credibility of rare variants [Ahituv et al., 2007; Madsen and Browning, 2009; Ng et al., 2009; Price et al., 2010; Pritchard, 2001]. These relationships can be easily incorporated into our Bayesian model by modifying the prior means for variants. By doing so, our approach has the additional advantage of accounting for uncertainties about these relationships in the hierarchical modeling.

There are several ways in which our method may be extended. First, for simplicity, we have not considered the issue how to determine which variants to be combined. The approach proposed by Li and Leal [2008] that pools variants below a fixed allele-frequency threshold

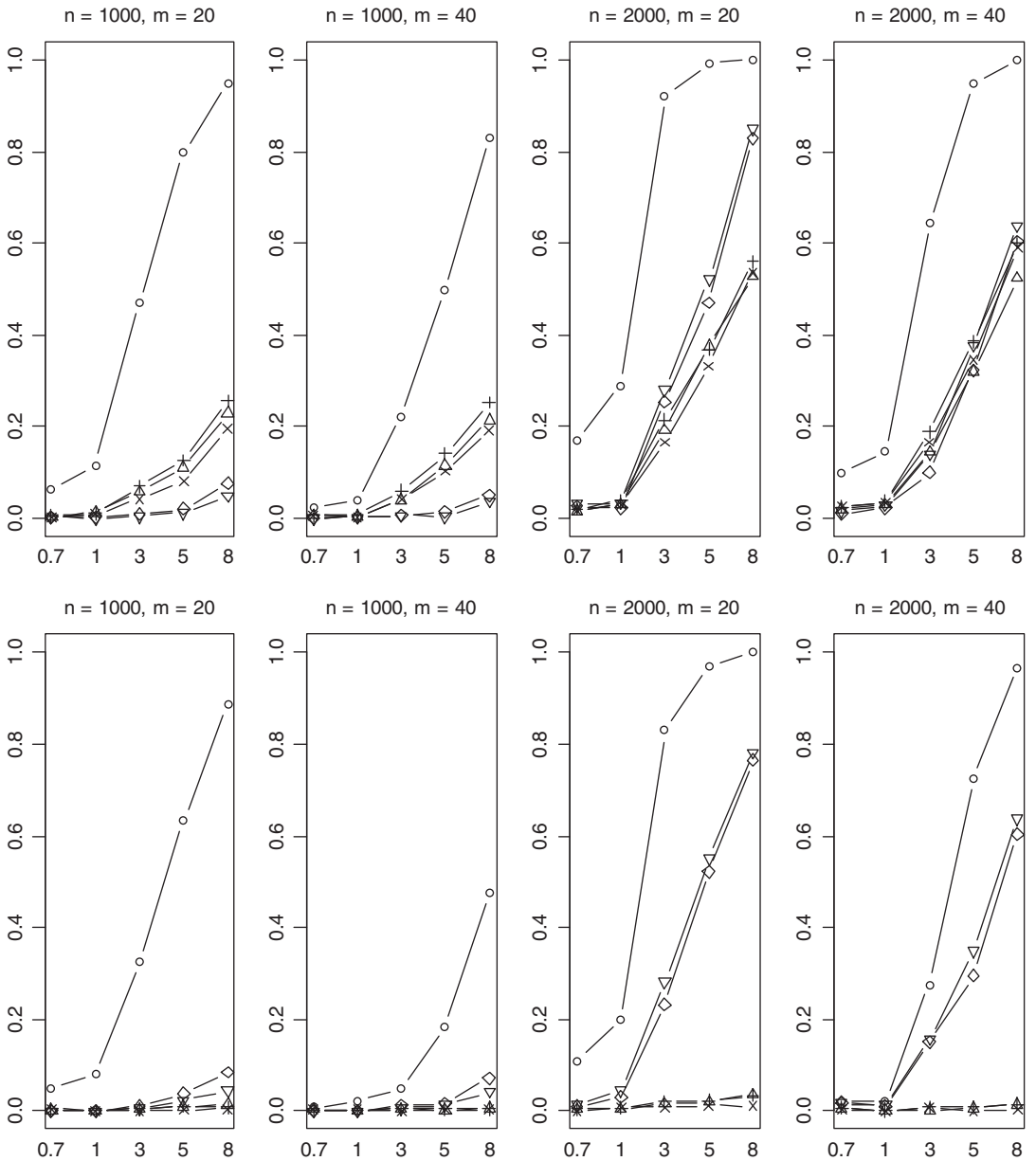

Fig. 8. Empirical powers of the proposed method (○), Collapsing (△), Simple-Sum (+), Weighted-Sum (×), All-Variants (◇), and Single-Variant (▽) methods at significance level of $\alpha = 0.001$. n and m represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. The top (bottom) pattern shows the analyses that 30% (50%) of the functional variants affect phenotype in the opposite direction.

(say, 1%) and separately models other variants can be easily applied to our model. A recently proposed method uses a variable allele-frequency threshold, which also can be incorporated into our model [Price et al., 2010]. Second, we have focused on rare variants in a gene or region, but complex diseases are usually influenced by multiple genes and environmental factors and their interactions. Our hierarchical model can be easily extended to include environmental factors as covariates and jointly analyze all rare variants in multiple genes using a separate genetic score for each gene. In principle, we can extend the proposed model to include gene-environment and gene-gene interactions by defining an overall coefficient and a genetic score for each interaction. However, it would be interesting to investigate statistical power for detecting interactions in analysis of rare variants. Third, rare variants tend to have occurred more recently and therefore population stratification should be adequately controlled when analyzing rare variants [Eichler et al., 2010]. We can infer population substructure from sufficient data and then incorporate them into our model.

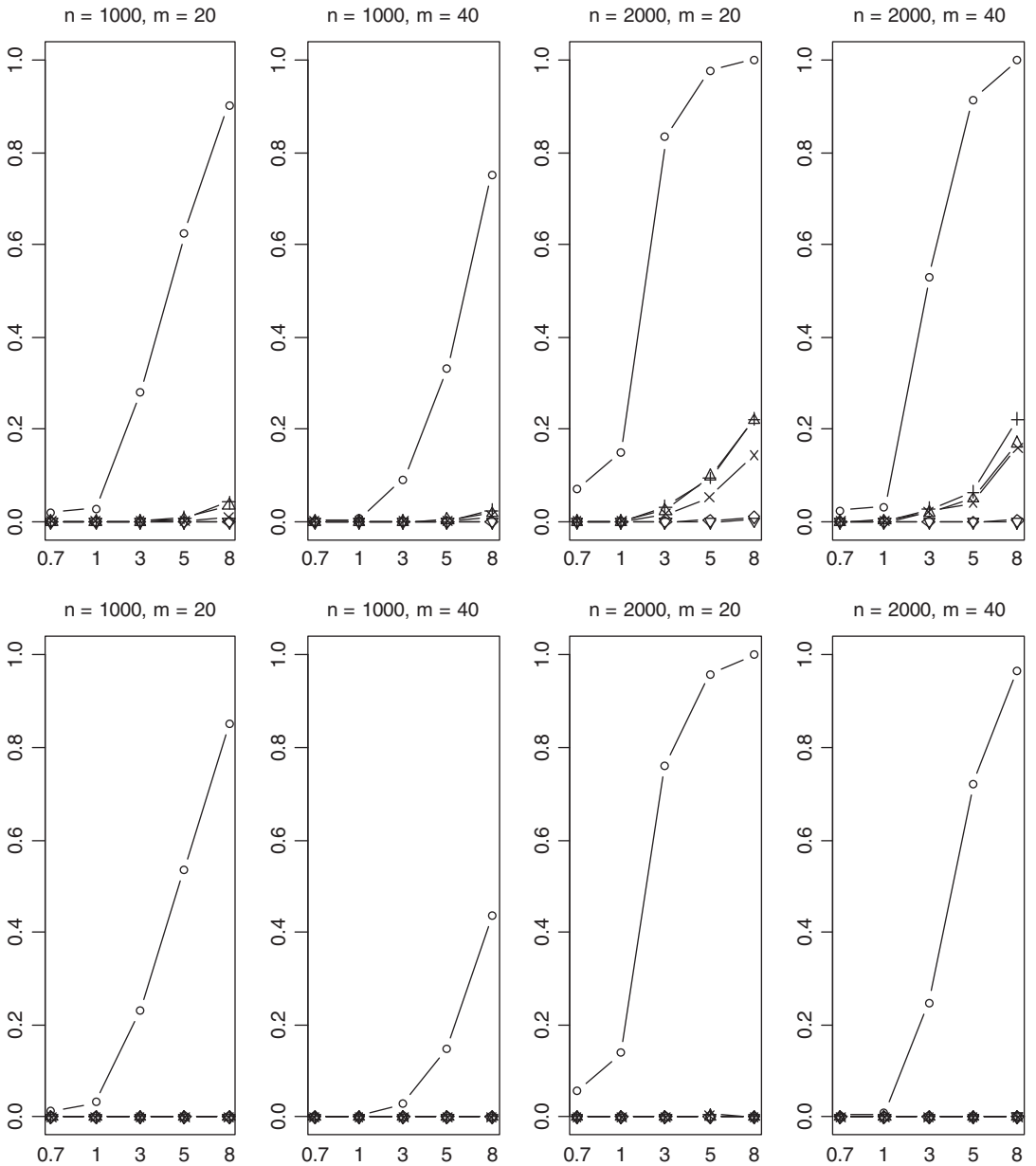

Fig. 9. Empirical powers of the proposed method (○), Collapsing (△), Simple-Sum (+), Weighted-Sum (×), All-Variants (◇), and Single-Variant (∇) methods at a significance level of $\alpha = 2.5 \times 10^{-6}$. *n* and *m* represent the numbers of individuals and rare variants, respectively. The total liability heritabilities are 0.7, 1, 3, 5, or 8%. The top (bottom) pattern shows the analyses that 30% (50%) of the functional variants affect phenotype in the opposite direction.

ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health (NIH) Grants 2R01GM069430-06 and GM077490 to N. Y. and the NIH Grant R00 RR024163 to D. Z. N. Y. and D. Z. designed the statistical models and simulation studies together. N. Y. implemented the method and developed the software. Both authors contributed to the writing of the manuscript. The authors declare no competing financial interests.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA. 2007. Medical sequencing at the extremes of human body mass. Am J Hum Genet 80:779–791.

Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, Rawstorne E, Colley J, Moskvina V, Frye C, Sampson JR, Wenstrup R, Scholl T, Cheadle JP. 2008. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res 68:358–363.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872.

Cohen JC, Boerwinkle E, Mosley Jr TH, Hobbs HH. 2006. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med 354:1264–1272.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450.

Gelman A. 2004. Parameterization and Bayesian modeling. J Am Stat Assoc 99:537–545.

Gelman A, Hill J. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York: Cambridge University Press.

Gelman A, Jakulin A, Pittau MG, Su YS. 2008. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2:1360–1383.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42–54.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40:592–599.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387–389.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: common disease-common variant…or not? Hum Mol Genet 11:2417–2423.

Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC. 2007. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39:513–516.

Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC. 2009. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119:70–79.

Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19: 212–219.

Wray NR, Yang J, Goddard ME, Visscher PM. 2010. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6:e1000864.

Yi N, Banerjee S. 2009. Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181:1101–1113.

Yi N, Kaklamani VG, Pasche B. 2010. Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. Ann Hum Genet, DOI: 10.1111/j.1469-1809.